A framework for integrating heterogeneous sporadic knowledge sources into automatic speech recognition

Stefan Ziegler, Guillaume Gravier

A Framework for Integrating Heterogeneous Sporadic Knowledge Sources into Automatic Speech Recognition

Stefan Ziegler, Guillaume Gravier

CNRS-IRISA, Campus de Beaulieu, 35042 Rennes, France

firstname.lastname@irisa.fr

Abstract

Heterogeneous knowledge sources that model speech only at certain time frames are difficult to incorporate into speech recognition, given standard multimodal fusion techniques. In this work, we present a new framework for the integration of this sporadic knowledge into standard HMM-based ASR. In a first step, each knowledge source is mapped onto a logarithmic score by using a sigmoid transfer function. Theses scores are then combined with the standard acoustic models by weighted linear combination. Speech recognition experiments with broad phonetic knowledge sources on a broadcast news transcription task show improved recognition results, given knowledge that provides complementary information for the ASR system.

Index Terms: multimodal fusion, landmark-driven ASR, event-based speech recognition

1. Introduction

Multimedia data in the form of broadcasts, podcasts as well as audio-visual content present difficult challenges for state-of-the art hidden Markov model (HMM) based automatic speech recognition (ASR), since ASR systems are still sensitive towards unseen speaking styles and changes in acoustic conditions. To improve acoustic modeling of HMM-based ASR, many studies advocate the incorporation of complementary knowledge sources into standard ASR to achieve improved recognition accuracy or robustness. Examples of such complementary knowledge sources are phonetic models, that aim at exploiting different features and modeling techniques motivated by phonological studies, to build reliable and sometimes highly specialized detectors for phonetic classes [1, 2, 3, 4]. Another example is audio-visual ASR, where, if available, the visual modality is added to the existing acoustic information, to benefit from the fact that acoustically similar speech classes might correspond to very different visual counterparts (visemes), that are reliable to detect [5]. While it has often been argued that it is desirable for each knowledge source to rely on individual features and modeling techniques, the common architecture of state-of-the-art ASR has become a bottleneck for seamlessly integrating heterogeneous knowledge into speech recognition. Consequently, external knowledge sources often rely on rather homogeneous standard modeling techniques, like frame-based Gaussian mixture models, that are integrated with conventional feature or decision fusion techniques inside the given architecture of HMM-based ASR.

In this paper, we present a new framework for integrating heterogeneous sporadic knowledge sources into HMM-based ASR, with the term sporadic referring to the fact that each knowledge is only defined at certain time frames, often referred to as *events* (e.g., [1, 6]) or *landmarks* (e.g., [7, 8]). Indeed, many acoustic or visual cues for phonetic events or visemes are naturally modeled as a sequence of discrete events, rather than continous values, which makes their integration into ASR very difficult, given common multimodal fusion techniques. In our framework, integration of theses knowledge sources into standard HMM-based ASR is performed in two steps: First, we map each knowledge source onto a logarithmic score, using a sigmoid transfer function. This allows the integration of knowledge sources of different scaling, that appear asynchronously and do model arbitrary phonemic classes. In a second step, the obtained scores are combined with the acoustic scores of standard HMM-based ASR using weighted linear combination. These modified acoustic scores are integrated into the Viterbi decoding of the first pass of a large vocabulary ASR system.

In audio-visual ASR, continuous visual knowledge is often integrated into ASR via feature-fusion, i.e., concatenating audio and visual features to train refined acoustic models [9]. This approach is also used for the integration of a burst onset landmark detector in [2]. Decision fusion at the frame level using GMMs and HMMs by weighted linear combination of log-likelihood scores is used for integration of phonetic information in [10] and for visual information in [11]. Phonetic knowledge is also integrated into ASR during the rescoring step of multi-pass ASR [3, 7]. Landmark-based phonetic models have been used inside alternative probabilistic ASR frameworks [12] and in [1] statistical-post processing of sporadic phonetic landmarks resulted in improved detection accuracy.

In the following section we will present our framework in detail, before presenting speech recognition experiments using broad phonetic knowledge sources. The paper will conclude with an outlook on future work.

2. Integration of sporadic knowledge into ASR

Given a speech utterance with t frames, we consider a sporadic knowledge source k to be a function $x_k(t)$, with $x_k(t)$ being defined only for n_k frames $\mathcal{T}_{x_k} = \{t_1, \ldots, t_{n_k}\}$. Each source is the result of an external system specialized in detecting a given set of phonemes $\mathcal{S}_k$, which is a subset of the complete set of phonemes (including non-speech symbols) $\mathcal{P}$, with $\mathcal{S}_k \subset \mathcal{P}$. To integrate this knowledge into triphone-based ASR systems, the phonemes in $\mathcal{S}_k$ have to be mapped to the corresponding states $\mathcal{I}_k$, which is equally a subset of the complete search space $\mathcal{I}$ (see Figure 2). While the range of $x_k(t)$ is arbitrary for each source k, for example one source could provide a probability from 0 to 1, while another source might correspond to a score in the range from $-\infty$ to $+\infty$ or $-\infty$ to 0, we assume a clear correlation between $x_k(t)$ and $\mathcal{S}_k$. Assuming positive correlation, low values for $x_k(t)$ are supposed to signal poor confidence in

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

the presence of $\mathcal{S}_k$ at t, while high values have a very low error rate, with a more or less sharp transition in-between.

To illustrate such external knowledge sources, we use the example of integrating phonetic landmark detectors into HMM-based ASR. Landmark detection usually consists of two steps (see for example [13]). First, the system detects potential locations for speech events (landmarks), before acoustic cues in vicinity of these landmarks are evaluated to estimate the probability of one or several phonetic classes for each landmark. For example, vowels can be detected by local maxima in the first formant frequency and evaluation of additional features around this landmark can specify the type of vowel. An additional detector might provide landmarks signaling the presence of plosives, by detecting abrupt changes in the signal and studying several cues, like voice onset time or energy of the burst around this point (see for example [14]). It is obvious, while the detection of vowels and plosives can be highly specialized for each phonetic class, both classes are only defined at very specific locations $\mathcal{T}_{x_k}$. Furthermore the landmarks for vowels and plosives will be attached with a confidence estimate $x_k(t)$ that cannot be compared with each other, since each class uses different classification algorithms and features.

With $x(t)$ not being defined for most t, sporadic knowledge can avoid to model parts of speech with high uncertainty about the acoustic content, which is a major advantage compared to HMM-based acoustic modeling. While heterogeneity, i.e., the fact that the ranges of each $x_k(t)$ are very different from each other, could be overcome by normalization, the sporadic nature of knowledge sources makes common fusion at the feature or decision level not feasible any more, since k knowledge sources cannot be mapped onto a k-dimensional vector at each frame t (see Figure 3).

In the following, we present a general framework for the integration of k knowledge sources into the Viterbi decoding of a HMM-based ASR system. Given k sources $x_k(t)$, two steps are necessary from raw knowledge to knowledge-driven ASR. First, we map each source x_k onto a log-likelihood score $\log s_k$, given a sigmoid transfer function, which parameters are estimated using cross-entropy as the objective function. In the second step, these knowledge sources are integrated into the ASR system using a weighted linear combination of the obtained scores $\log s_k$ and the acoustic scores of the ASR system.

2.1. Weighted linear combination of k knowledge sources

Given k knowledge sources, our goal is to modify the acoustic score $s(i,t)$ for state $i \in \mathcal{I}$ at frame t according to weighted linear combination of the log-likelihoods of k knowledge sources $\log s_k(i,t)$ and the unweighted log-likelihood of the acoustic model $\log s_{asr}(i,t)$, given the weights w_k:

$$\log s(i,t) = \log s_{asr}(i,t) + \sum_k w_k \log s_k(i,t) \qquad (1)$$

With $\log s_k(i,t) \geq 0$ and $w_k \geq 0$, each source k enhances states $i \in \mathcal{I}_k$ that are associated with the phonemes in set $\mathcal{S}_k$ (see Figure 2). Evidently, $\log s_k(i,t) = 0$ for all states $i \notin \mathcal{I}_k$ and for all frames t for which the source k is not defined with $t \notin \mathcal{T}_{x_k}$. All states $i \in \mathcal{I}_k$ share the same likelihood-score $\log s_k(i,t)$, to which we will refer to as $\log s_k(t)$.

The next section describes how to map $x_k(t)$ onto $\log s_k(t)$ for each source, before we discuss determining w_k.

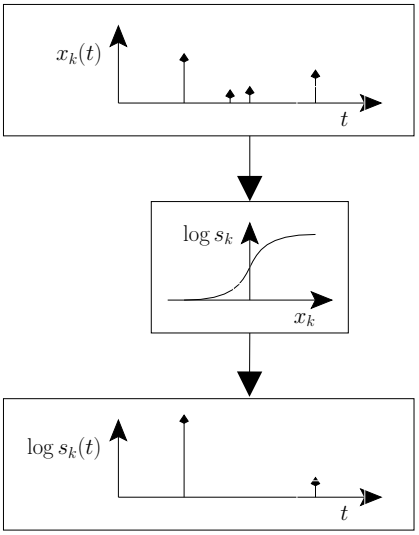

Figure 1: Mapping a sporadic knowledge source $x_k(t)$ onto $\log s_k(t)$.

2.2. Mapping of detection functions onto knowledge scores

Intuitively, $\log s_k(t)$ should maximize the scores added to the correct path, i.e., the scores added to frames t where the correct phoneme actually is a member of $\mathcal{S}_k$, but minimize the error it will introduce into the system by enhancing the wrong path. Therefore, our mapping function should result in $\log s_k(t) = 0$ for low values of $x_k(t)$, but grow according to the confidence that higher values of x_k will correctly indicate $\mathcal{S}_k$. This desired behavior can be obtained by a sigmoid function with:

$$\log s_k\,(t) = \frac{\gamma_k}{1 + \exp\left(-\alpha_k \cdot x_k\,(t) + \beta_k\right)},\ \forall t \in \mathcal{T}_{x_k} \qquad (2)$$

α_k determines the steepness of the slope of the sigmoid, β_k shifts the sigmoid to its optimal working point and γ_k is a scaling factor. For example, if a knowledge source k provides a very reliable knowledge above a certain score β_k, γ_k will be a high value reflecting the confidence in the correctness of $\log s_k(t)$ and a high α_k changes the transfer function from a smooth transition to a step-function-like behavior. Equation 2 maps noisy, unreliable values onto values very close to zero and rounding those values to a limited precision results in $\log s_k(t) = 0$. Since $\log s_k(t) = 0$ for all $t \notin \mathcal{T}_{x_k}$, $\log s_k(t)$ is effectively a sparse vector and we refer to its non-zero frames as $\mathcal{T}_{s_k}$.

To find the optimal α_k, β_k and γ_k, we maximize the cross-entropy $c_{ce}(t)$ between $\log s_k(t)$ and the correct solution $y_k(t)$ at each frame:

$$c_{ce}\,(t) = y_k(t)\frac{\log p_k(t)}{N_{k,1}} + (1 - y_k(t))\,\frac{\log\left(1 - p_k(t)\right)}{N_{k,0}} \qquad (3)$$

$y_k(t)$ is a binary vector with $y_k(t) = 1$ if $\mathcal{S}_k$ is correct at frame t and $y_k(t) = 0$ if not. $y_k(t)$ is derived from the forced alignment of the correct utterance using our baseline ASR system. $p_k(t)$ reflects the probability that knowledge source k is present at frame t. Since some knowledge sources might have a skewed distribution, we normalize $p_k(t)$ by the number of frames $N_{k,1}$ that are in $\mathcal{T}_{x_k}$ for which $y_k(t) = 1$ and respectively $N_{k,0}$ for which $y_k(t) = 0$.

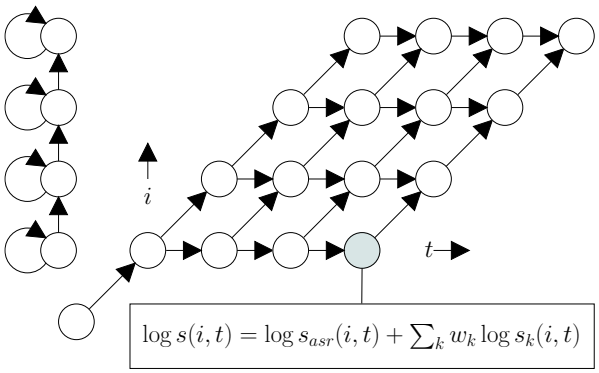

Figure 2: Integration of knowledge into the speech decoding. Arrows correspond to the transition probabilities, while the nodes represent the acoustic scores $\log s(i,t)$. The modified computation of $\log s(i,t)$ is displayed for one node highlighted in grey.

Given the log-likelihood scores of two complementary classes s_k and $\overline{s}_k$, we use the softmax function to estimate $p_k(t)$ according to:

$$p_k(t) = \frac{\exp(\log s_k(t))}{\exp(\log s_k(t)) + \exp(\log \overline{s}_k(t))} \quad (4)$$

As a consequence of the facts that all knowledge sources might model only a subset of $\mathcal{P}$ and sporadic knowledge results in asynchronous landmarks, there is no score $\log \overline{s}_k(t)$ estimating the *absence* of knowledge source k at frame t. Consequently, this *anti-score* $\log \overline{s}_k(t)$ always equals 0:

$$\log \overline{s}_k(t) = 0, \ \forall t \quad (5)$$

The final optimization problem consists in finding the parameters α_k, β_k and γ_k that maximize $c_{ce}(t)$ for all frames of the training data:

$$F_{ce,k}(\alpha_k, \beta_k, \gamma_k; x_k, y_k) = \sum_{t \in \mathcal{T}_{x_k}} c_{ce}(t) \quad (6)$$

2.3. Estimation of the combination weights

While the optimized knowledge sources $\log s_k(t)$ might achieve low error rates according to Equation 6, it has yet to be determined if this source represents *complementary* knowledge to the acoustic models of the ASR system. Therefore, we use discriminative training to determine the weight w_k for each source k, that adjusts the contribution of source k to the overall acoustic score according to Equation 1.

Estimating the weights w_k of a linear combination of log-likelihoods is a well studied problem and several discrimination criteria have been proposed in the literature [15, 11, 10]. In this paper we use the frame-based maximum mutual information (MMI) between correct alignment and n competing hypothesis according to:

$$c_{mmi}(t) = \log s(u(t),t)) - \log \sum_n \exp(\log s(\hat{u}_n(t))) \quad (7)$$

$u(t)$ is the state sequence obtained by force aligning the correct solution of an utterance, while $\hat{u}_n(t)$ corresponds to the alignment of the n-th hypothesis contained in the n-best output of the ASR system. By maximizing the MMI, the correct hypothesis will become more likely, while at the same time the competing hypothesis that do not correspond to the correct path at frame t will become less likely. In this work, we use only the best hypothesis as a competing alternative to the correct path, so that $n = 1$, which turns the MMI criterion into corrective training (see [15]). The optimization problem consists then in finding the weights w_k that maximize $c_{mmi}(t)$ over all frames in $\mathcal{T} = \bigcap_k \mathcal{T}_{s_k}$:

$$F_{mmi}(w; u, \hat{u}, \log s) = \sum_{t \in \mathcal{T}} c_{mmi}(t) \quad (8)$$

3. Experiments

The corpus used in the experiments corresponds to radio broadcast news in the French language from the ESTER2 campaign [16]. The ESTER2 dataset contains broadcast shows with speech in studio environments (RFI), but also difficult tasks like debates (Inter) or speech with strong accents (radio TVME and radio Africa 1). Since we need the correct hypothesis to generate the correct state sequences $u(t)$ and the aligned n-best recognition hypothesis $\hat{u}_n(t)$, we discard every sentence containing out-of-vocabulary words during training and testing. During testing, this allows us to assure that finding the correct path by modifying the acoustic scores during the decoding is not prevented by missing vocabulary. Additionally, we discard all telephone speech from the dataset. The estimation of the parameters α_k, β_k, γ_k and w_k are conducted on the ESTER2 development set, using only broadcasts shorter than 20 minutes, while final speech recognition experiments are conducted on the full ESTER2 test set. The speech recognizer used in this paper is a two-pass system, trained on the ESTER1 and ESTER2 training data. The first pass uses word-internal triphones with 32 Gaussians per state and a trigram language model. The second pass relies on 4-grams and cross-word triphone models. In this paper, we integrate knowledge only in the first pass of our ASR system to generate improved word graphs for rescoring.

3.1. Phonetic knowledge sources and baseline ASR system

In the experiments, we use broad phonetic classes (BPCs) as knowledge sources, obtained from the Gaussian mixture models of a Mel-frequency cepstral coefficients based monophone GMM classifier. We derive 6 detection functions $x_k(t)$ for the BPCs vowels, nasals, approximants, fricatives, plosives and a non-speech class. Each BPC at frame t is first scored with the maximum score among all phonemes of this BPC, before we perform normalization at each frame t by taking the logarithmic sum of exponentials for each source k to obtain 6 continuous detection functions. After smoothing we convert these 6 functions into $k = 6$ sporadic knowledge sources $x_k(t)$ by simple picking the local maxima for each detection function (see Figure 3). Since the monophone models were trained on the same training data like our acoustic models, it is unlikely that they actually will provide complementary information to the ASR system. To experiment with more informative knowledge sources, we additionally create oracle knowledge by adding a bias to the correct BPC at each frame t before performing normalization. We refer to this knowledge sources as BPC-oracle-$bias$, with $bias$ being the scalar added to the correct BPC. While this knowledge does not represent homogeneous knowledge in the sense that it incorporates different modeling and training frameworks, we discuss the influence of multiplicative and additive scaling of

knowledge	WER [dev]	WER [test]
baseline	28.0	31.8
BPC-0	28.0	31.8
BPC-oracle-2	27.7	31.6
BPC-oracle-3	27.4	31.3
BPC-oracle-4	26.8	31.0

Table 1: Word error rates of 4 different broad phonetic knowledge sources and the baseline ASR system on the ESTER2 development and test set.

each $x_k(t)$ in section 3.5.

3.2. Optimization

Given k knowledge sources $x_k(t)$, we have to optimize two objective functions to obtain the parameters α_k, β_k and γ_k for each knowledge source individually and the weights w_k jointly. We use L-BFGS-B minimization implemented in pythons scipy library for both objective functions, with the constraints $\alpha_k > 0$ and $\gamma_k \geq 0$ for Equation 6 and $w_k \geq 0$ for Equation 8. The gradients of the objective functions are in both cases calculated using the symbolic differentiation implemented in the Theano package [17].

For both objective functions, we could achieve fast convergence by carefully choosing initial values for both optimization problems. The scaling factor α_k should be proportional to the variance of $x_k(t)$, while the median of $x_k(t)$ is a good starting point for β_k. For Equation 8, we started with the same value w_k for all knowledge sources k, by choosing the uniform weight which maximized Equation 8. This lead to Equation 6 needing about 20 iterations to converge, while Equation 8 converged already after very few iterations. Though we maximized the weights w_k globally, instead of using gradient descent, we did not observe problems concerning convergence or overfitting.

3.3. Speech Recognition Experiments

After optimizing the mapping from $x_k(t)$ to $\log s_k(t)$ for all sources k and estimating the weights w_k on the development set, speech recognition experiments were performed for *BPC-0*, *BPC-oracle-2*, *BPC-oracle-3* and *BPC-oracle-4*. Table 1 shows the word-error-rates (WER) on the ESTER2 development and test-set along with the WER of the baseline. As expected, *BPC-0* did not provide any new information for the ASR system and obtained $w_k = 0$ for all BPCs except for the non-speech class. Consequently this led to no improvement in WER. For the oracle BPCs, the WER decreases with increasing the bias of the knowledge source. For all cases, the improvement on the development set is higher than on the test set, as often observed in discriminative training.

3.4. Evaluation of knowledge sources

Table 2 displays two criteria evaluating the quality of $x_k(t)$ and $\log s_k(t)$ for the BPCs of three different experiments. AUC (area under the curve) is a performance measurement derived from the ROC curve (receiver operator characteristic) and equal to the probability that a classifier will rank a randomly selected true BPC higher than a randomly selected false BPC. We use the AUC to give an indication of the quality of the raw knowledge source $x_k(t)$. Additionally, for every knowledge source k, we calculate a misclassification cost (MI), related to the mutual information criterion MMI in Equation 7, by calculating the average score added at each frame $\mathcal{T}_{s_k}$, with weighting every correct frame by 1 and every incorrect frame by -1. This results in a negative value if a knowledge source introduces more errors into the decoding than it enhances the correct path.

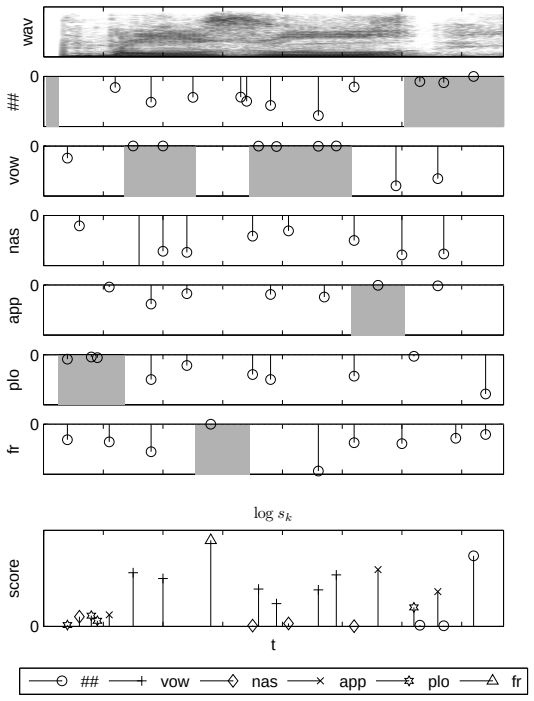

Figure 3: Spectrogram of the French word *Bonjour*, uttered at the beginning of a broadcast show, followed by six sporadic broad phonetic knowledge sources $x_k(t)$ (*BPC-oracle2*) including non-speech (##) and the obtained log likelihoods $\log s_k(t)$ at the bottom. All $x_k(t)$ are normalized, so that 0 represents the maximum value. The correct sequence of BPCs is marked in grey.

$$MI(k, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (2y_k(t) - 1) \log s_k(t) \qquad (9)$$

$|\mathcal{T}|$ corresponds to the cardinality of the frames $\mathcal{T}$ used to calculate $MI(k, \mathcal{T})$. Both measures are shown on all available frames $\mathcal{T}_{x_k}$ for AUC and $\mathcal{T}_{s_k}$ for MI. Additionally, they are calculated only on those frames $\mathcal{T}_k^*$ where the correct BPC of the true alignment $u(t)$ differs from the BPC in $\hat{u}_n(t)$.

Since the acoustic score of the standard ASR system is not modified (see Equation 1), we expect an improvement of the WER only if a knowledge source is able to correctly enhance most of the frames that are not already correctly aligned in the best recognition hypothesis. Indeed, it can be seen that *BPC-0*, while performing relatively well on all frames $\mathcal{T}$, has a below random AUC, with $AUC < 0.5$, for all BPCs except silence for $\mathcal{T}^*$. For those BPCs MI is negative, which means these knowledge sources make it less likely for the decoder to find the correct path at frames $\mathcal{T}^*$. Consequently, discriminative training resulted in $w_k = 0$ for all BPCs except silence, to prevent the ASR system from degrading. In general, evaluating the errors of a knowledge source without taking the output of the speech recognizer into account might be misleading. Only

BPCs		$\mathcal{T}$	##	vow	nas	plo	fri	app
BPC 0	AUC	$\mathcal{T}_k$	0.84	0.90	0.95	0.93	0.96	0.83
		$\mathcal{T}_k^*$	0.41	0.43	0.37	0.35	0.34	0.46
	MI	$\mathcal{T}_k$	0.9	0.6	2.1	1.7	2.5	0.8
		$\mathcal{T}_k^*$	0	-0.1	-0.5	-0.5	-0.8	-0.1
BPC oracle 2	AUC	$\mathcal{T}_k$	0.94	0.96	0.98	0.98	0.99	0.93
		$\mathcal{T}_k^*$	0.67	0.63	0.59	0.61	0.53	0.70
	MI	$\mathcal{T}_k$	1.9	1.0	3.3	3.0	3.1	1.9
		$\mathcal{T}_k^*$	0.7	0.4	0.5	0.6	-0.2	0.6
BPC oracle 4	AUC	$\mathcal{T}_k$	0.98	0.99	0.99	0.99	0.99	0.98
		$\mathcal{T}_k^*$	0.87	0.81	0.80	0.83	0.73	0.88
	MI	$\mathcal{T}_k$	3.0	1.7	4.6	4.2	3.6	3.2
		$\mathcal{T}_k^*$	1.9	1.3	2.1	2.1	0.8	2.0

Table 2: AUC for $x_k(t)$ and MI for $\log s_k(t)$ given different knowledge sources and their broad phonetic classes silence and non-speech (##), vowels, nasals, plosives, fricatives and approximants. $\mathcal{T}_k$ corresponds either to $\mathcal{T}_{x_k}$ for AUC or $\mathcal{T}_{s_k}$ for MI.

when knowledge sources $x_k(t)$ achieve above random AUC on $\mathcal{T}^*$, MI tends to turn positive and the source contributes to improving the WER, as it is the case for *BPC-oracle-2* and *BPC-oracle-4*.

3.5. Heterogeneous knowledge

The previous broad phonetic knowledge sources were obtained using homogeneous monophone GMM classifier and thus did not represent a collection of heterogeneous knowledge sources. Assuming heterogeneous knowledge will change $x_k(t)$ into $x'_k(t)$ by multiplicative and additive scaling with $x'_k(t) = a_k x_k(t) + b_k$, it is evident that given our proposed sigmoid transfer function, this scaling can be reversed by estimating the corresponding α_k and β_k. To avoid the problem of finding an individual initialization for α_k and β_k to optimize objective function 6 for each knowledge source, we recommend to perform a simple normalization, for example mean and variance normalization for each knowledge source $x_k(t)$. All of our experiments showed, that given proper initialization for α_k and β_k, $\log s_k(t)$ and consequently $MI(k, \mathcal{T})$ was similar for different multiplicative and additive scaling factors.

One advantage of our presented framework is the fact that it is able to deal with selected knowledge sources, that may not cover the complete set of phonemes $\mathcal{P}$. This allows to design individual detectors for each phonemic group $\mathcal{S}_k$, without forcing to model the whole set $\mathcal{P}$. Table 3 shows the same speech recognition experiments as in section 3.3, but with the reduced set of BPCs vowels, nasals and plosives. It can be seen that the WER increases compared to using the complete range of BPCs and the overall impact of the provided knowledge sources is reduced. This is expected, since the broader the external knowledge sources become, the less impact they will have onto the speech decoding, even if a knowledge source inserts only few errors into the decoding.

4. Future Work

Our presented framework showed promising results given different kinds of broad phonetic knowledge sources. Before concluding the paper we want to point out several directions for future research.

Knowledge sources: Our experiments showed, while the integration of rather broad speech landmarks into HMM-based ASR improves the recognition, these landmarks need to be accurate to be effective. Obviously, efforts have to be made to research on existing and new knowledge sources that provide sufficiently accurate landmarks. Furthermore, it is desirable to experiment with additional feature systems like distinctive features, or visual features like visemes.

knowledge	WER [dev]	WER [test]
baseline	28.0	31.8
BPC-oracle-2 (vow-nas-pl)	27.6	31.8
BPC-oracle-3 (vow-nas-pl)	27.5	31.7
BPC-oracle-4 (vow-nas-pl)	27.3	31.5

Table 3: Word error rates of 3 different broad phonetic knowledge sources, using only the BPCs vowels, nasals and plosives each time.

Objective functions: While the sigmoid transfer function in connection with the cross-entropy criterion in Equation 6, as well as the MMI criterion for discriminative training provided good results, one might consider additional transfer functions and training criteria.

State dependent weights and context dependency: One disadvantage of the presented approach is the fact, that it does not include state or phoneme-dependent weights $w_{i,k}$ for Equation 1. Enhancing states that are not in $\mathcal{I}_k$ for a knowledge source k might help to reduce the error introduced into the decoding, since this might take into account common phonetic confusions, like it is the case for vowels and approximants. Additionally, the speech recognition system could be modified to accommodate for a weight w_{asr} that scales $\log s_{asr}(i,t)$ in Equation 1 to improve the discriminative training criterion.

Given phonetic landmarks, as employed in this paper, the probability of a speech class $\mathcal{S}_k$ at t depends on the context, i.e., its preceding and subsequent landmarks. To address this context dependency, landmarks $x_k(t)$ could be rescored by additional models, that are trained on landmark sequences, like it has been proposed in [1].

Integration into multi-pass ASR: In the current implementation we only implemented knowledge-driven ASR in the first pass of our speech recognizer. To fully benefit from heterogeneous knowledge sources, integration into rescoring steps of multi-pass ASR systems is desirable.

5. Conclusions

The presented framework focused on the integration of heterogeneous and sporadic knowledge sources into HMM-based ASR. It allows the use of individual training and detection algorithms for each knowledge source, that can be developed independently from each other. Furthermore, it accounts for event or landmark based models of speech and does not require the re-training of existing acoustic models. We used a transfer function to map each knowledge source onto a logarithmic score, before the obtained values were combined with the acoustic scores by weighted linear combination.

While the knowledge sources that improved the WER in this paper corresponded to oracle knowledge, we conclude from our experiments that landmarks which achieve an above random detection performance on frames where the ASR-system aligns the wrong path are likely to improve the recognition performance of HMM-based ASR systems.

6. References

[1] A. Jansen and P. Niyogi, "Point process models for event-based speech recognition," *Speech Communication*, vol. 51, no. 12, pp. 1155–1168, 2009.

[2] C.-Y. Lin and H.-C. Wang, "Burst onset landmark detection and its application to speech recognition," *Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1253–1264, 2011.

[3] S.M. Siniscalchi and C.H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, no. 11, pp. 1139–1153, 2009.

[4] S. Ziegler, B. Ludusan, and G. Gravier, "Towards a new speech event detection approach for landmark-based speech recognition," in *2012 IEEE Workshop on Spoken Language Technology*, 2012.

[5] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in Visual and Audio-Visual Speech Processing*, vol. 22, pp. 23, 2004.

[6] A. Juneja, *Speech recognition based on phonetic features and acoustic landmarks*, Ph.D. thesis, University of Maryland, College Park, 2004.

[7] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, "Landmark-based speech recognition: report of the 2004 Johns Hopkins summer workshop," in *Proc. of ICASSP'05*, 2005, pp. 213–216.

[8] G. Gravier and D. Moraru, "Towards phonetically-driven hidden Markov models: can we incorporate phonetic landmarks in HMM-based ASR?," in *Proc. of NOLISP'07*, 2007, pp. 161–168.

[9] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual lvcsr," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, 2001, pp. 165–168.

[10] F. Metze, "Articulatory features for 'meeting' speech recognition," in *Proc. of INTERSPEECH-2006*, 2006.

[11] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and mce based hmm stream weight estimation for audio-visual asr," in *Proc. of ICASSP'02*, 2002.

[12] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Comput. Speech Lang.*, vol. 17, pp. 137–152, 2003.

[13] Amit Juneja and Carol Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 123, pp. 1154–1168, 2008.

[14] Sharlene A Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am.*, vol. 100, pp. 3417–3430, 1996.

[15] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, no. 3, pp. 287–310, 2001.

[16] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French broadcasts," in *Proc. of INTERSPEECH-2009*, 2009, pp. 1149–1152.

[17] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010, Oral Presentation.